Fermi National Accelerator Laboratory

FERMILAB-Conf-89/51
[SSC-206]

Summary Talk: Data Acquisition, Event Building, and On-Line Processing*

Irwin Gaines
Fermi National Accelerator Laboratory
P.O. Box 500, Batavia, Illinois 60510

March 1989

* Presented at the Workshop on Triggering and Data Acquisition for Experiments at the Superconducting Super Collider, Toronto, Canada, January 16-19, 1989.

Operated by Universities Research Association, Inc., under contract with the United States Department of Energy

Summary Talk:
Data Acquisition, Event Building, and On-Line Processing

Irwin Gaines
Fermilab Advanced Computer Program, Batavia, IL 60510

Our subgroup of working group 4 concerned itself with general architectural issues for SSC data acquisition. Fiber optic buses were described in two talks, and software and project management were discussed in a separate subgroup. These topics are covered in separate papers. A number of different issues were considered, with certain natural differences of opinion, and a list of projects needing further R&D is given at the end of this paper. The major conclusion, however, can be simply stated and was agreed to unanimously: that the SSC data acquisition systems can achieve much higher data rates than those assumed at previous workshops. On-line processor farms of 10^5 VAX equivalents (throughout this paper we will use VAX 11/780 units as a performance standard) are quite conceivable, and data recording rates of between 100 and 1000 Hz will be feasible.

ARCHITECTURE AND RATES

The architectural framework for the subgroup's discussions is shown in fig. 1. We assumed that data would be buffered on (or near) the detector in systems discussed by the front-end working group, and that a prompt trigger together with some data processing at the front end would reduce both the event rate and event size before the actual data acquisition system began to deal with the events. Our discussions were confined to the area below the dotted line in the figure, where a stream of digital data emerges from the front end systems.

The components of the DA system that we considered include the non-prompt triggers, data and control buses, event builders, high level language processing farms, and data recording ("tape"). Also, it is important to point out that when we give a rate capability for a particular component of the data acquisition system, that does not imply that we are required to run the system at that rate. In particular, it will be necessary to make choices, guided by physics, as to where the biggest payoff comes from making improvements in the rate capability. It might cost the same amount of dollars to double the "tape" writing speed as to increase the power in the processor farm by 20%, and we will need physics judgement to decide where to put our resources. Our discussions were primarily aimed at identifying the "brick walls" in the rate capabilities, beyond which we would require significantly new designs or unanticipated technological breakthroughs.

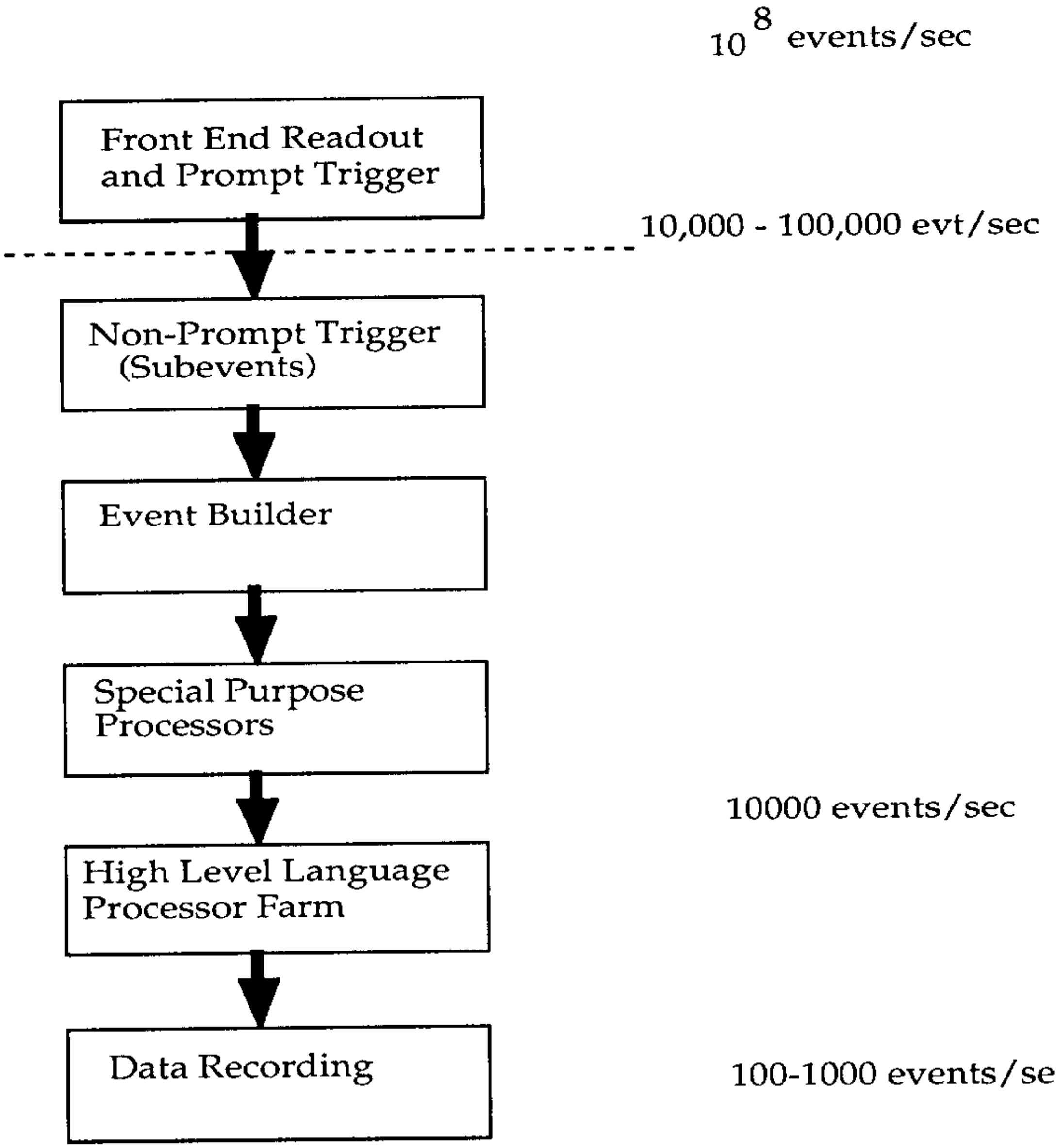
10^8 events/sec
Front End Readout and Prompt Trigger
10,000 - 100,000 evt/sec
Non-Prompt Trigger (Subevents)
Event Builder
Special Purpose Processors
10000 events/sec
High Level Language Processor Farm
Data Recording
100-1000 events/sec

Figure 1. Block diagram of data acquisition architecture.

With that in mind, we can summarize the rate capabilities as shown in the figure. We can take between 10^4 and 10^5 events/sec out of the front-end, with an event size of between 10^5 and 10^6 bytes. Special purpose non-prompt triggers could deal with these rates. The high level language processor farm can deal with 10^4 events/sec. Data can be recorded at somewhere between 100 and 1000 Hz, depending on how much data reduction is done in the DA system. The remainder of this paper will justify these numbers and summarize our discussions on several architectural issues.

EVENT BUILDING

Discussion of event builders considered the difference between "classical" event builders and newer ideas of using a switching network as an event builder (see fig. 2). It was generally agreed that the traditional designs, where one (or a few) event builders act as a funnel through which all the data must pass, were unacceptable for SSC rates. Schemes based on switching networks, on the other hand, have the advantage of being scalable to match the required rates and of being able to keep all the input and output pipelines busy at the same time. Mark Bowden presented one design based on a barrel shifter; other more complicated designs are also possible. Note that events can be processed and rejected both in the input pipelines (where the trigger processors work on sub-events) and in the output pipelines (where the processors can work on the entire event).

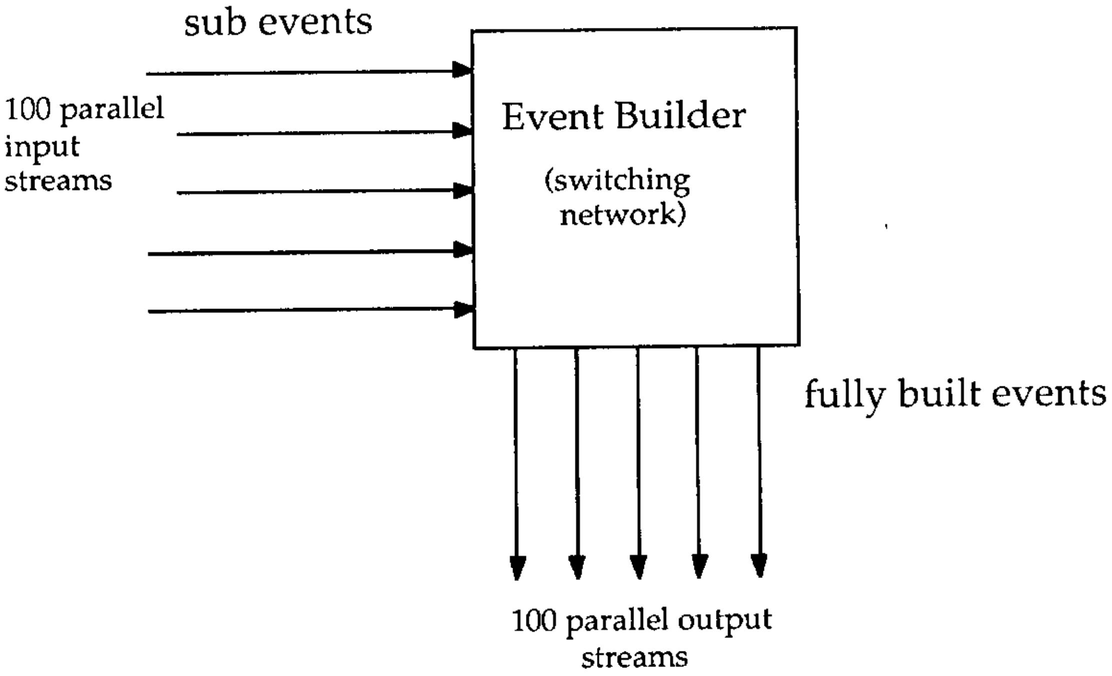

Figure 2. Event builder based on switching network.

At the highest rates contemplated, of 10^5 events/sec of size 1 MB each out of the front end, the input rate to the event builder will be 100 GBytes/sec. Assuming we have fiber optic data links capable of 1 GByte/sec, this would require a parallelism of 100 into and out of the event builder, which seems quite feasible.

TO BUS, OR NOT TO BUS

Discussion of buses centered not around which bus standard was more or less suitable to the SSC, but instead on whether or not we need a bus at all. There was general agreement on the need to separate out data and control paths, and on the need for some sort of diagnostic and control bus connecting most of the DA system. However, data transmission may likely be handled with a standard for a high-speed data link rather than a bus. Optical fiber technology appears capable of providing us with such a link that could work at rates of up to 1 GByte/sec. The separate control bus can run at much lower speeds.

DATA DRIVEN ARCHITECTURES

We discussed the relevance of data flow/data driven architectures to SSC DA systems. These ideas seem well matched to the needs for pipelines that are clearly evident in the front-end systems. Cited as potential advantages of such data driven architectures were simplicity, modularity, flexibility, and speed (since centralized control is unnecessary). Issues that were raised but not fully answered include the problems of decentralized control (maintaining synchronization across parallel pipelines, insuring no data loss between pipeline stages when the next stage is busy, etc.) and understanding how to propagate a fast reject signal through a pipeline.

SPECIAL VS. GENERAL PURPOSE PROCESSORS

Despite the enormous power that will be available in the form of general purpose processors, it was felt that there will continue to be a place for more specialized designs, for such tasks as data manipulations (pedestal suppression and rescaling) as well as segment and track finding and jet cluster finding. The potential utility of such special purpose devices is illustrated by two examples from CDF: cluster finding takes roughly a millisec per cluster on the VAX, but the special purpose hardware finds clusters at 200 nanosec per cluster, giving a performance of 5000 VAX equivalents; while a tracking task (finding all tracks with more than 3.5 GeV of transverse momentum) which takes over 100 millisec on the VAX is done in 2-4 microsec with limited resolution and 20-30 microsec at full resolution in the special purpose hardware, again giving a performance of better than 5000 VAX equivalents.

The input rate to the special purpose non-prompt triggers is expected to be between 10^4 and 10^5 events per second, depending on how much filtering is done by prompt triggers at the front end. The high-level language processor farm will be capable of accepting 10^4 events/sec. Thus, at the highest rates we will require at least a factor of 10 rejection from special purpose devices; while at the lowest rates we may simply perform data compression and preprocessing at this stage. Even at these lower rates, however, we may choose to reject events at this stage so as to put less of a burden on the processor farm.

Regardless of the amount of rejection needed from special purpose devices, as much use as possible should be made of commercial programmable devices, including DSPs and ASICs with RISC processor cores, to avoid the proliferation of processors that are only understood by a few experts.

HIGH LEVEL LANGUAGE PROCESSOR FARMS

The availability of powerful RISC processors allows us to consider farms of order 10^5 VAX equivalents. This will permit 10 VAX seconds of processing per event at an input rate of 10^4 events/sec, or more processing if there is additional reduction at the special purpose processor stage. Such a system could be constructed out of 500 processor boards, each with 4 50 MIP processors. The price of the processor boards should be under $5000 by the mid 1990's, making the total cost of the farm of order $2.5 million. Note that each processor board need only process 20 events/sec, so the data bandwidth needed into each board is only 20 MBytes/sec. This stage needs to provide a further event rejection of between 10 and 100 to match the "tape" writing capabilities.

EVENT COMPACTION

It would be useful if the size of the events could be substantially reduced before "tape" writing so as to reduce the required bandwidth for data recording. One might even hope to write out only data summary tapes (DSTs) rather than raw data given the large amounts of on-line processing power that will be available. Two factors mitigate against this.

First, it is unlikely that final calibration constants will be known to sufficient accuracy as the data is being taken to allow the raw data to be thrown away. At best, an automated production line could hope to determine constants within a few hours after the data is taken, allowing the DSTs to be produced shortly after the raw data is recorded. However, this will not relieve the DA system from the need for recording all the raw data, even if just temporarily.

More importantly, it appears that the event size on the DSTs is in fact larger than the size of the raw event! CDF, for example, has raw event sizes averaging 120 kBytes, while the DST event size is 250 kBytes. Only at the mini-DST stage is the event size reduced (to about 25kB/event). Thus, even if we had up-to-date calibrations it is unlikely that we can reduce the event size by additional on-line processing.

Nevertheless, the raw event size of 1 MByte at the SSC includes significant contributions from noise and from out-of-time events. On-line processing in the farm might be able to eliminate much of this non-event data, reducing the demands on the "tape" system. More detailed detector simulation (as well as greater understanding of the readout schemes) is required to know how much of a help this will be.

"TAPE' WRITING SPEEDS AND OFF-LINE ANALYSIS

Thus, we need to contemplate recording events of close to 1 MByte in size. Even so, it was generally agreed that the "tape" system can easily cope with 100 events/sec (100 MB/sec), with the possibilities of going another factor of 10 faster if necessary. Certainly we will be writing in parallel, with multiple streams of events of different trigger types. A parallel 100 MB/sec system could be built with today's technology for a cost of well under $1 million, and we can expect to do significantly better by the time the SSC turns on. It is still too early to decide the exact media we will record data on; high density tape cartridges, 8 mm video cartridges, and optical disks are all candidates. We will almost certainly need a "juke-box" like technology whatever the media is to minimize required user/operator intervention.

Before we consider recording 100-1000 events/sec, we should make sure that we will have off-line computing resources available to analyze the events. This is not a problem, however. A processor farm of the same scale as described above (10^5 VAX equivalents) can deliver 1000 VAX seconds per event at 100 Hz data taking rates or 100 VAX seconds at 1000 Hz, which should be sufficient.

OPEN QUESTIONS NEEDING R & D

Finally, we list some issues which were not resolved but were identified as needing additional work:

1) Simulation of DA systems--we will need sophisticated simulations of data and control flow at an early stage of the design of SSC data acquisition systems.
2) Fault tolerance and redundancy--It was generally felt that we do not need totally redundant readout paths for all data, but we do need to understand carefully just how much can go wrong before we no longer are taking useful data. Simulation will help a lot here.
3) Use of expert systems for fault diagnosis and debugging.
4) Playback of (Monte Carlo or real) data through the DA system.
5) Buses--what should be the standard for high speed data links (presumably some form of fiber optics);
   what should be used for the diagnostic/control network;
   what use can be made of data driven communications protocols;
   are any of the emerging new bus standards (Futurebus, SCI, etc) or any existing bus standards of any relevance to SSC.
6) We need a standard suite of benchmarks for processor evaluation.
7) Can we use high performance RISC processor cores in an application specific integrated circuit (ASIC); how will we interconnect such powerful special processors.